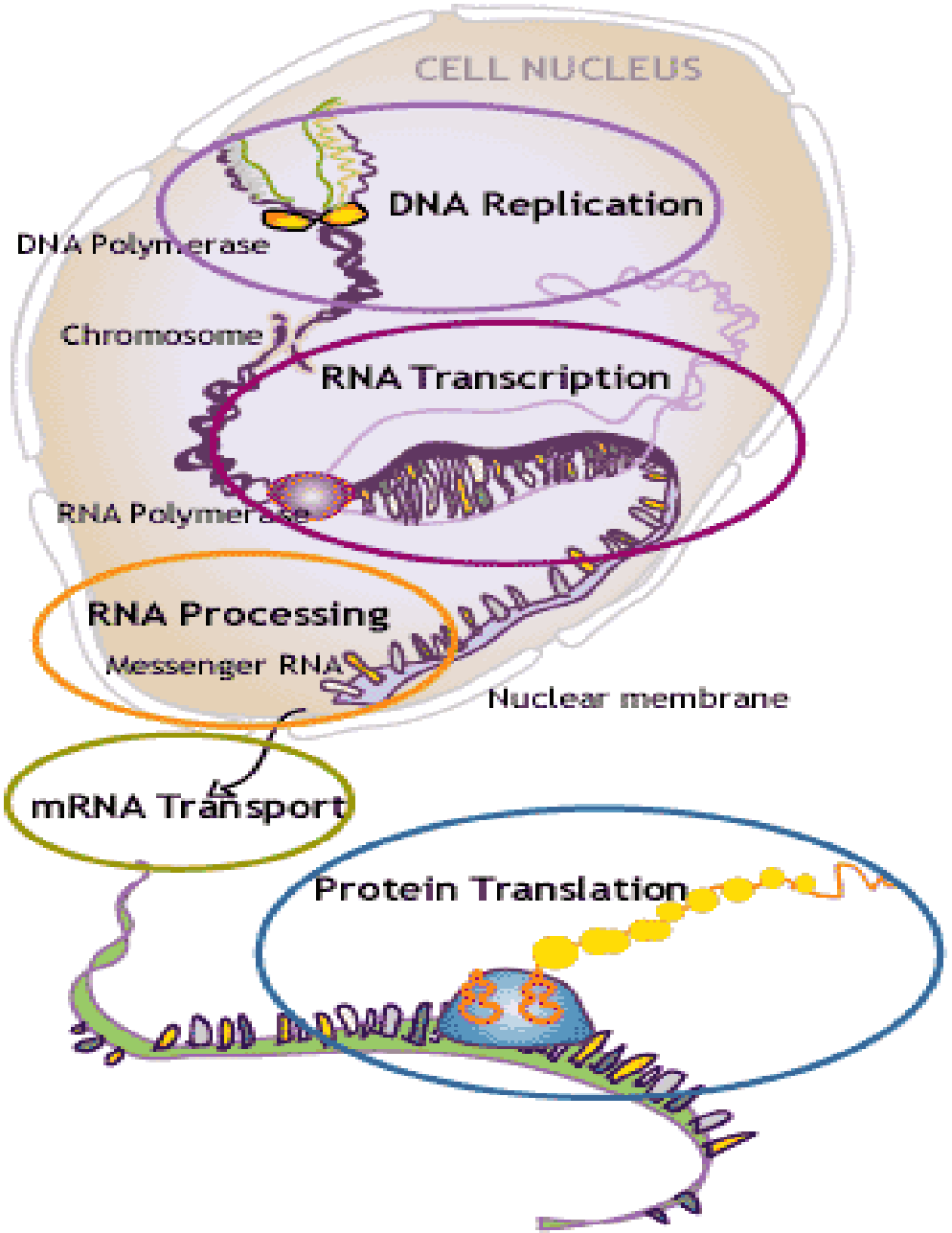
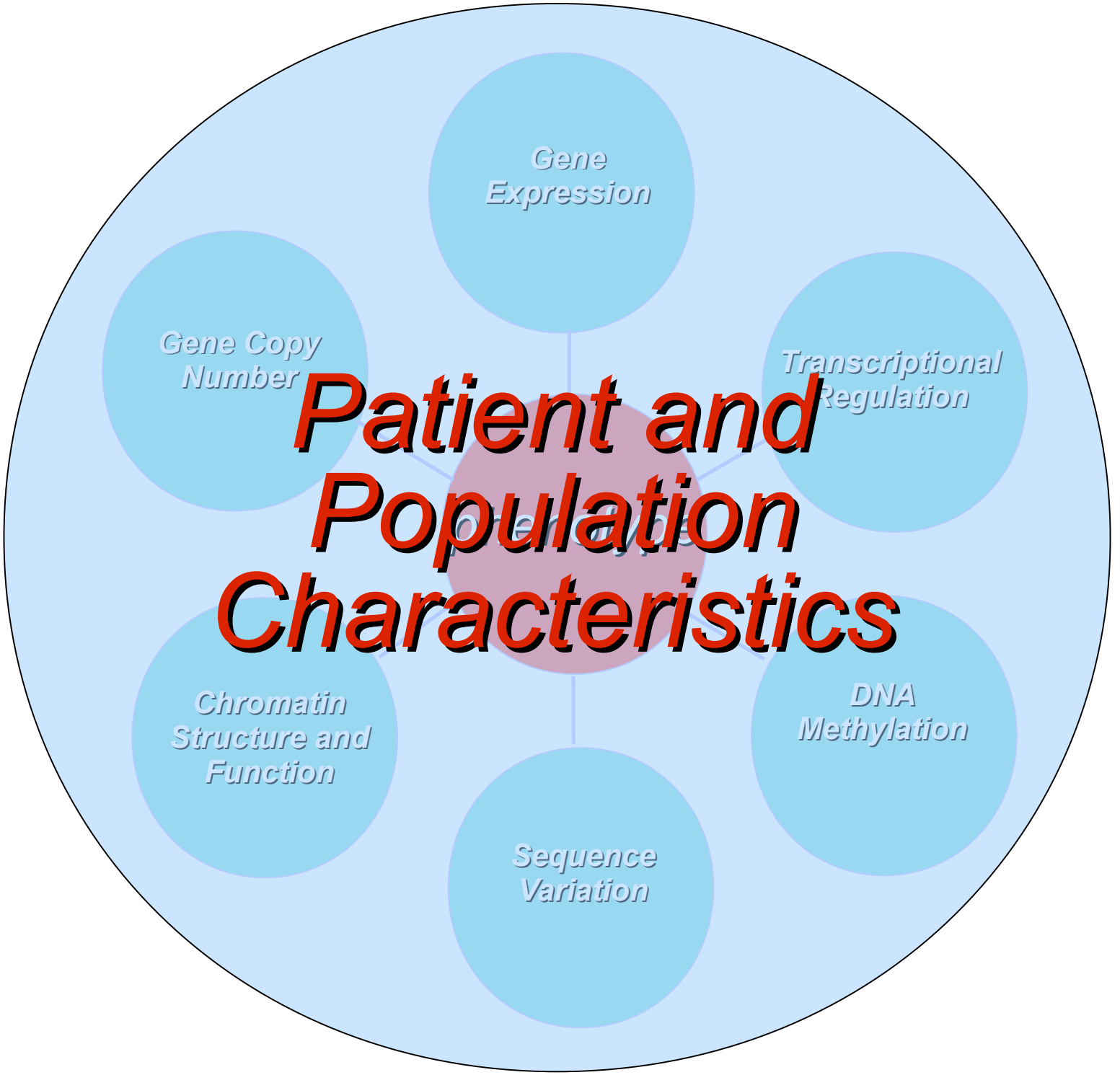
The Central Dogma




Gene Expression
Gene Copy Number
Transcriptional Regulation
Phenotype
Chromatin Structure and Function
DNA Methylation
Sequence Variation
Patient and Population Characteristics

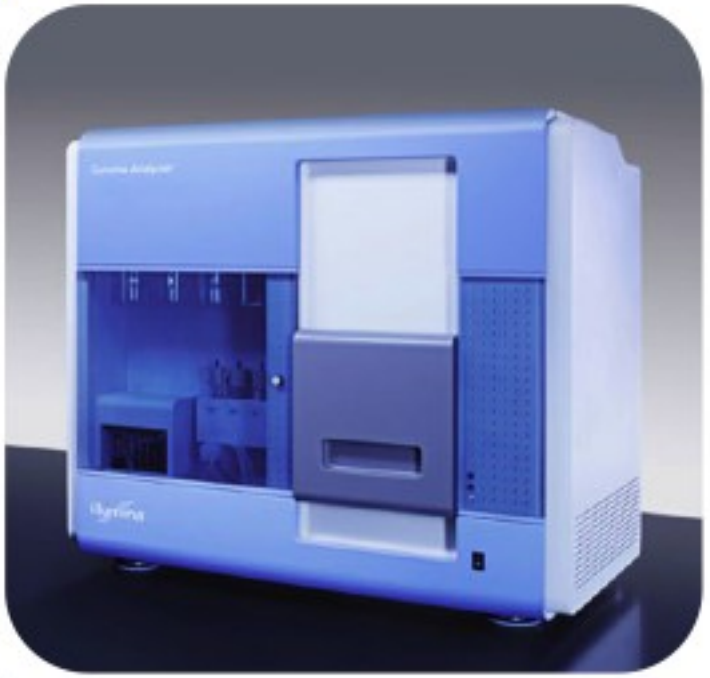
Sequencing

- Why use hybridization, which is just a measure of sequences, correct?
  - Sequencing is costly, time- and labor-intensive, and inefficient
- Next-generation sequencing technology changes the equation such that sequencing can, for some applications, be more efficient, cheaper, and less time- and labor-intensive than hybridization-based methods like microarrays

Illumina GA2

Setting the Standard in "Now Generation" Sequencing

Industry Leading Ease of Use…



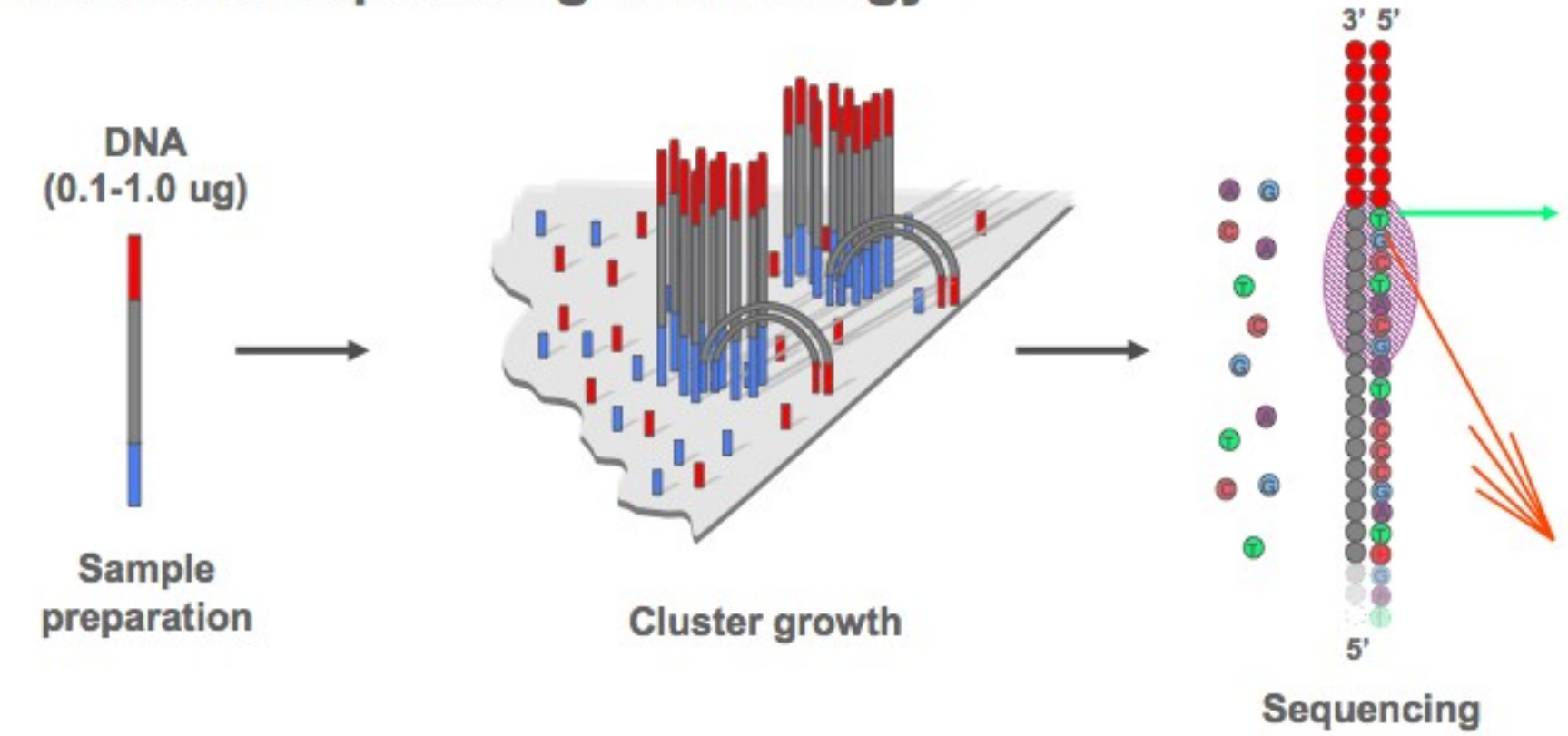
- Highest accuracy
- Highest percentage of perfect reads
- 1.5Gb+ single read in 2.5 days
- 3.0Gb+ paired read in 5 days
- 600MB per day today (minimum performance specification)
  - Supports future expansion
- Lowest operating cost
- >250 instruments installed

…with Applications Flexibility

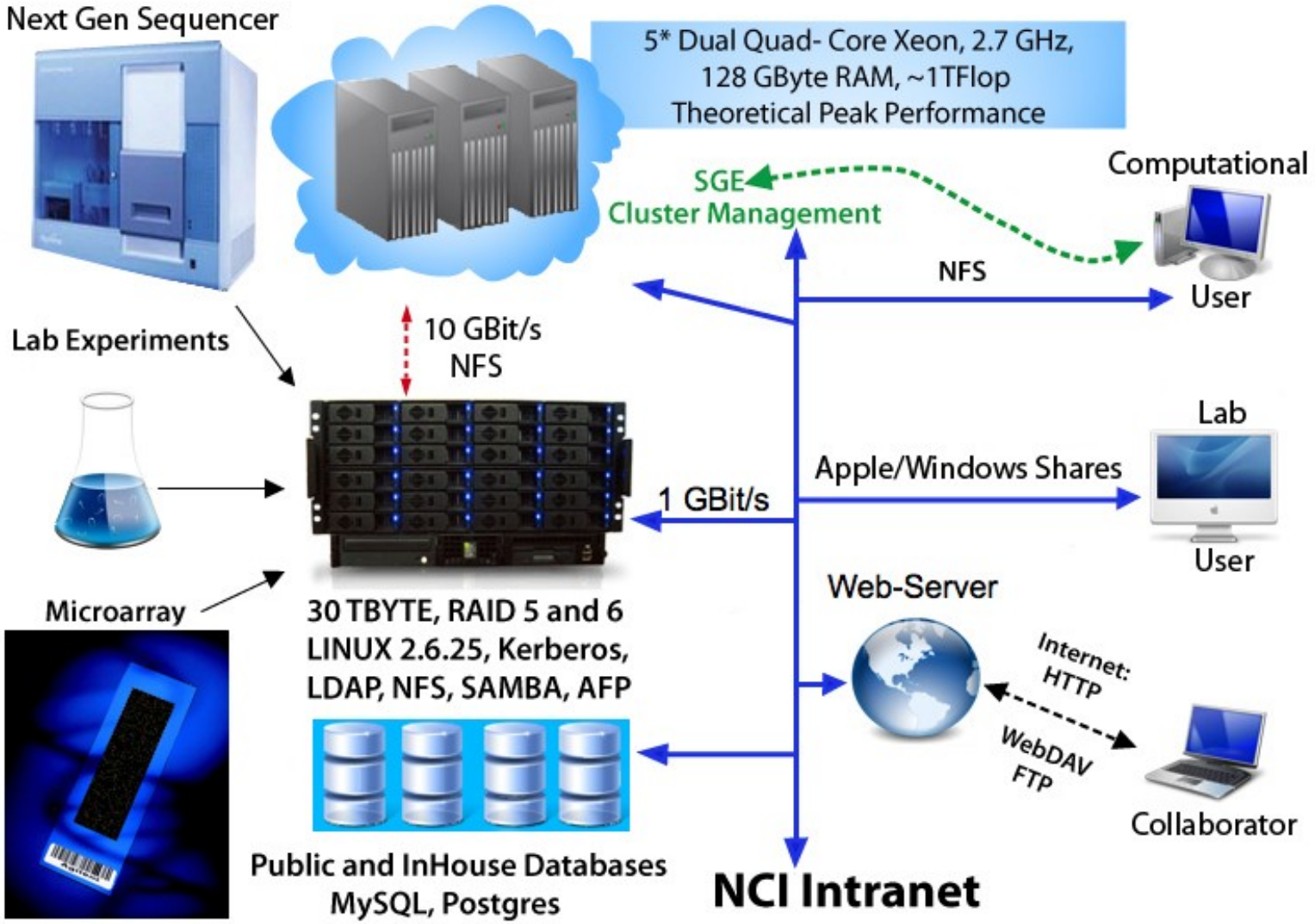
illumina

illumina

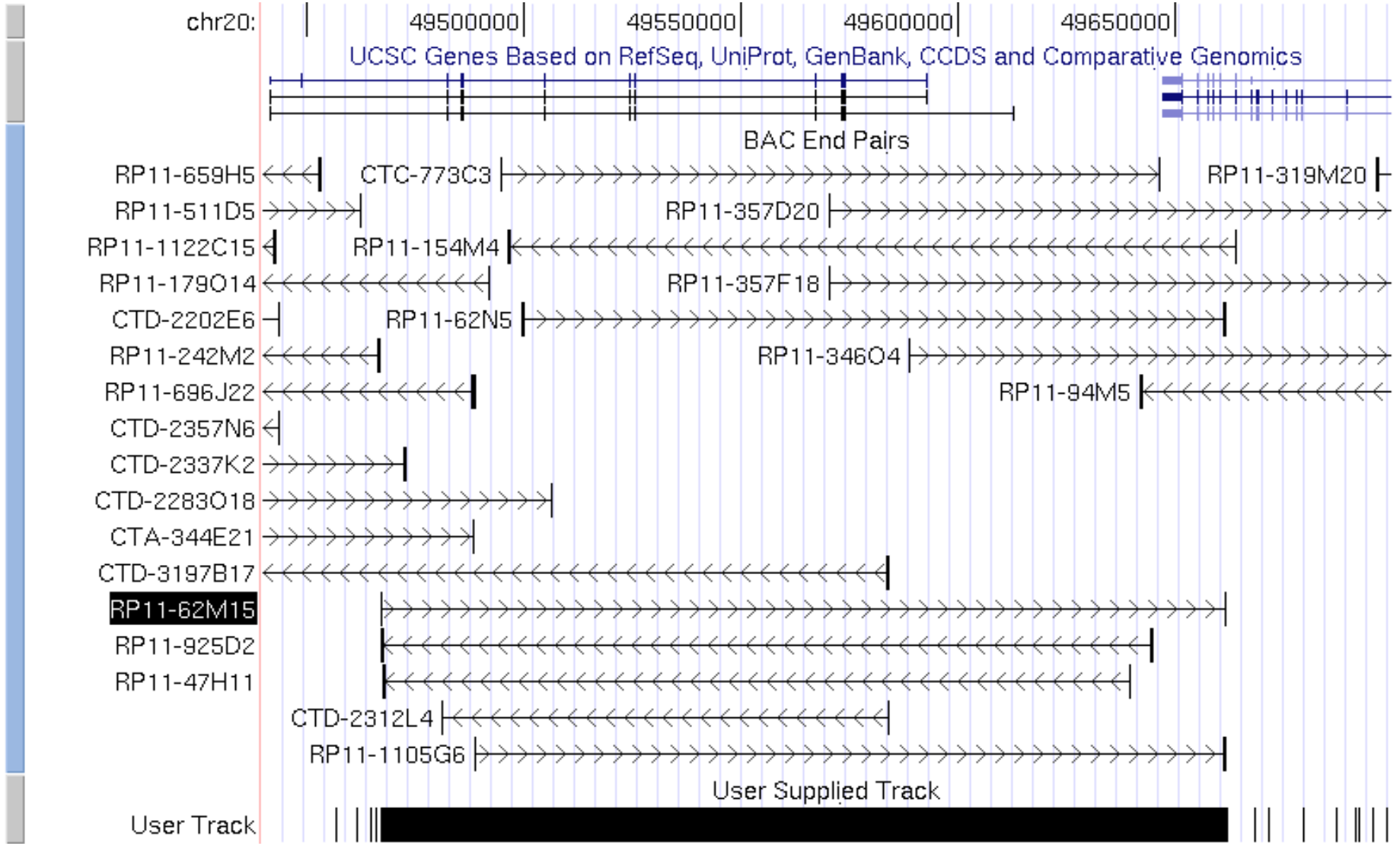
Illumina Sequencing Technology



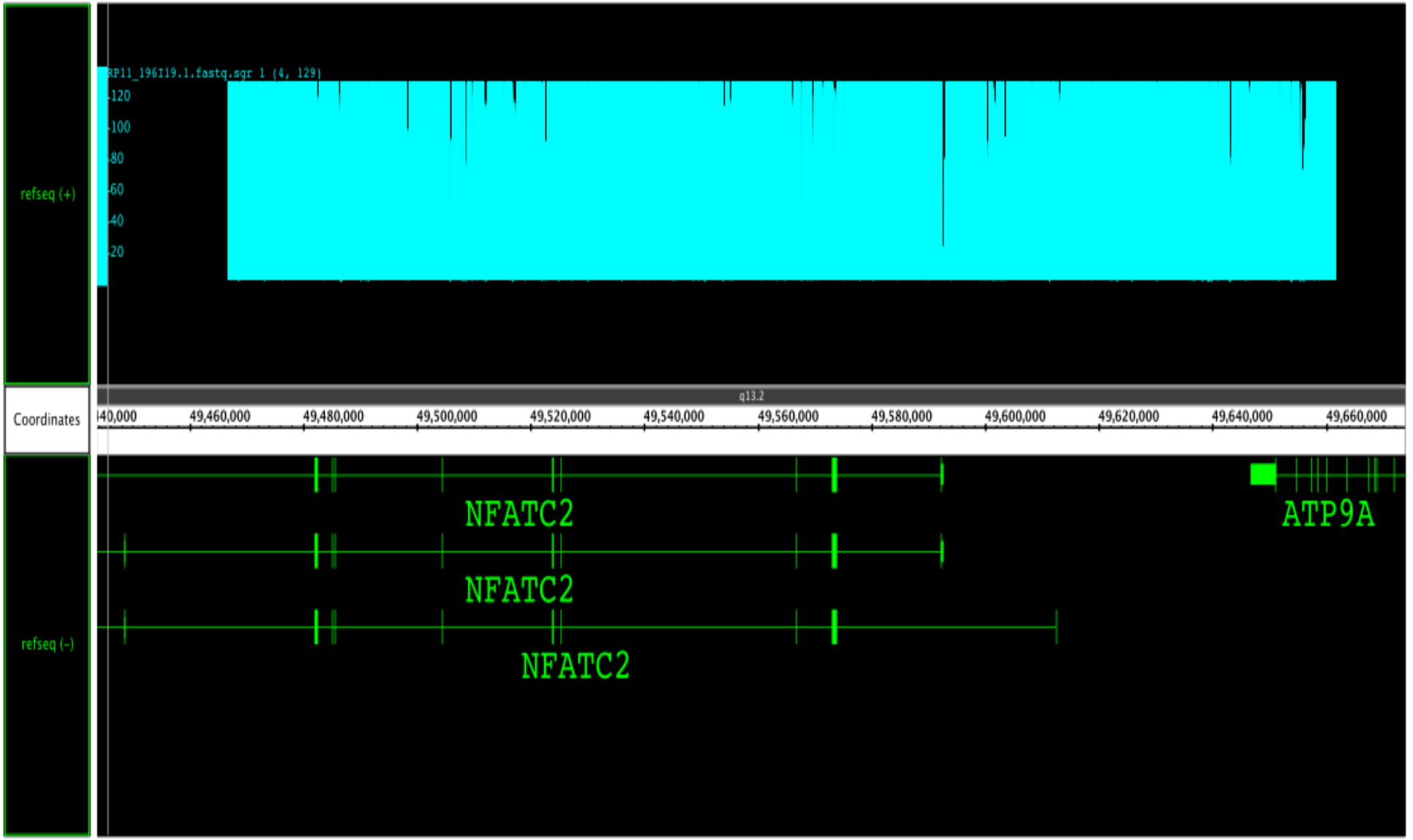
illumina


Next Gen Sequencer
5* Dual Quad- Core Xeon, 2.7 GHz,
128 GByte RAM, ~1TFlop
Theoretical Peak Performance
SGE
Cluster Management
Computational
NFS
User
Lab Experiments
10 GBit/s
NFS
Lab
Apple/Windows Shares
1 GBit/s
User
Web-Server
Microarray
30 TBYTE, RAID 5 and 6
LINUX 2.6.25, Kerberos,
LDAP, NFS, SAMBA, AFP
Internet:
HTTP
WebDAV
FTP
Collaborator
Public and InHouse Databases
MySQL, Postgres
NCI Intranet

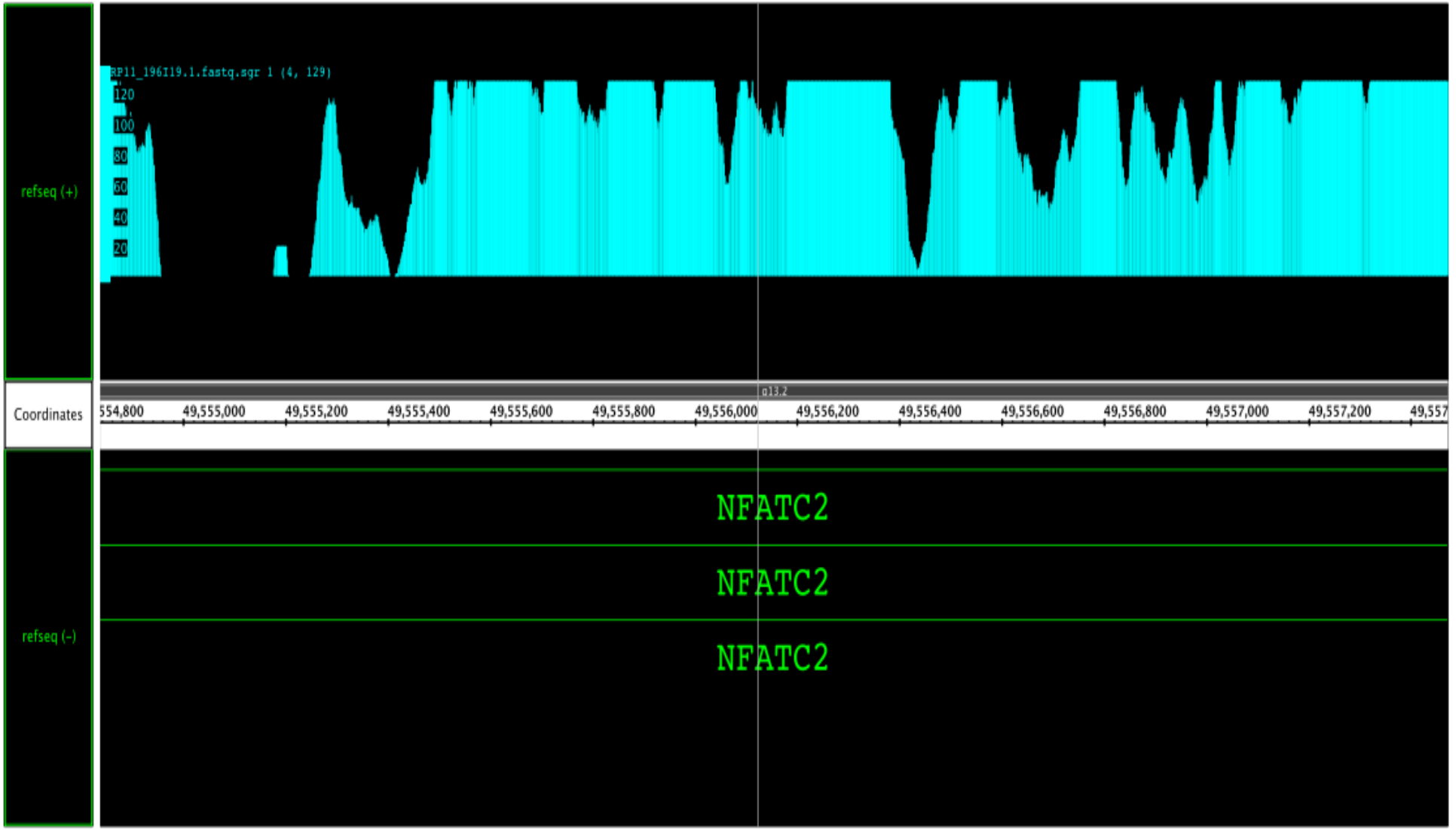
Overview of BAC in the Genome



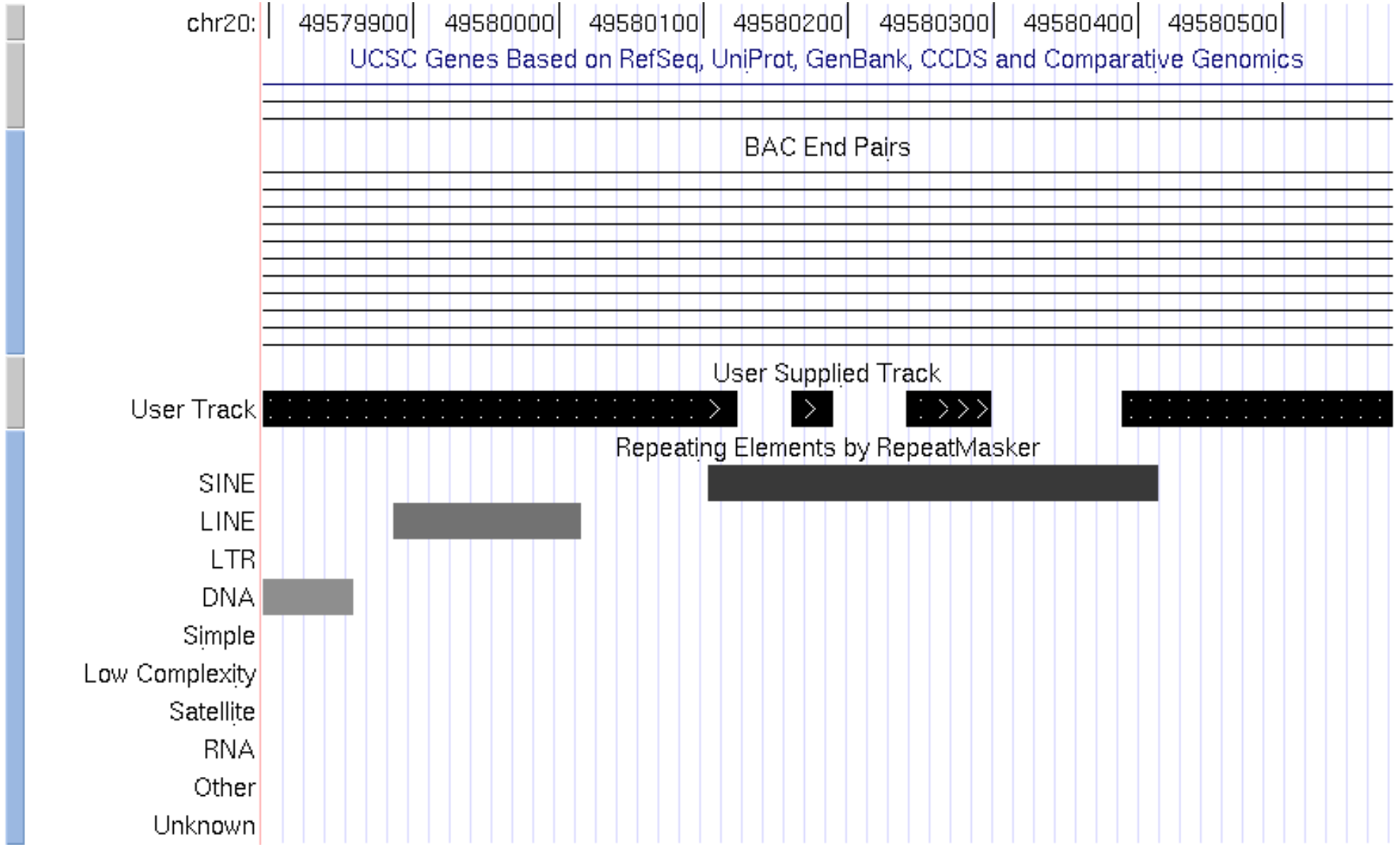
Sequencing a BAC



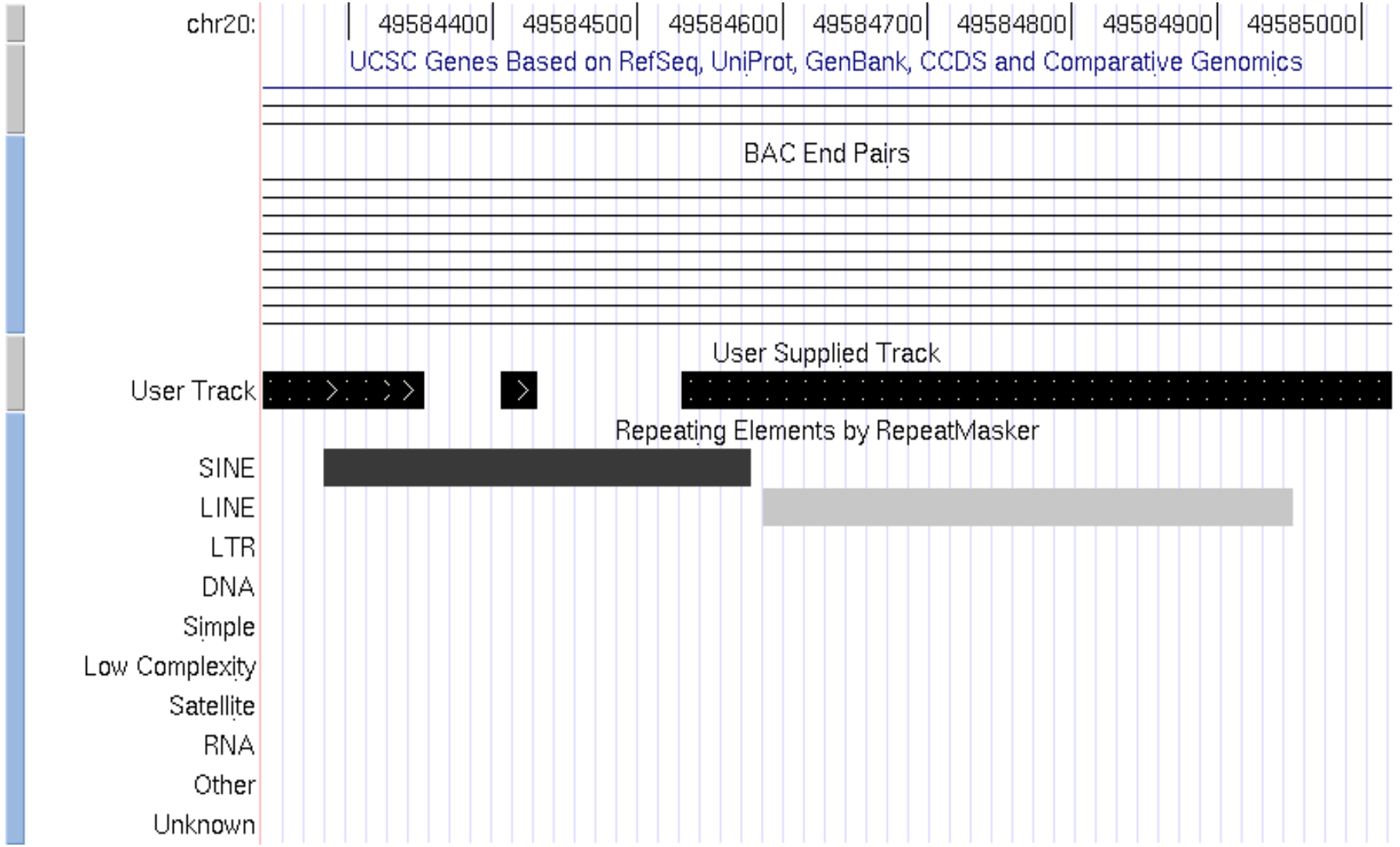
Sequence Coverage



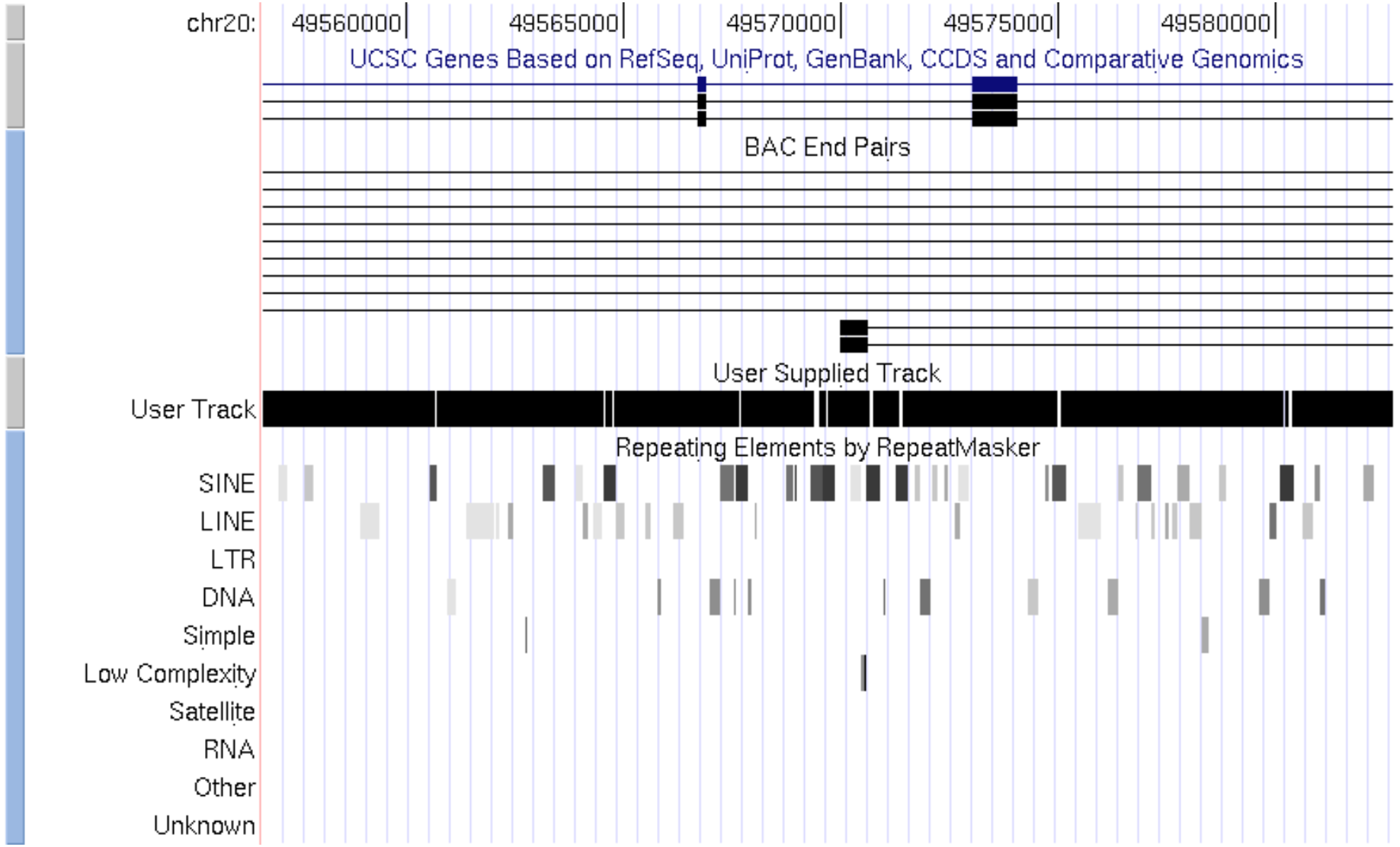
Repeats


chr20:
49579900
49580000
49580100
49580200
49580300
49580400
49580500
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BAC End Pairs
User Supplied Track
User Track
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

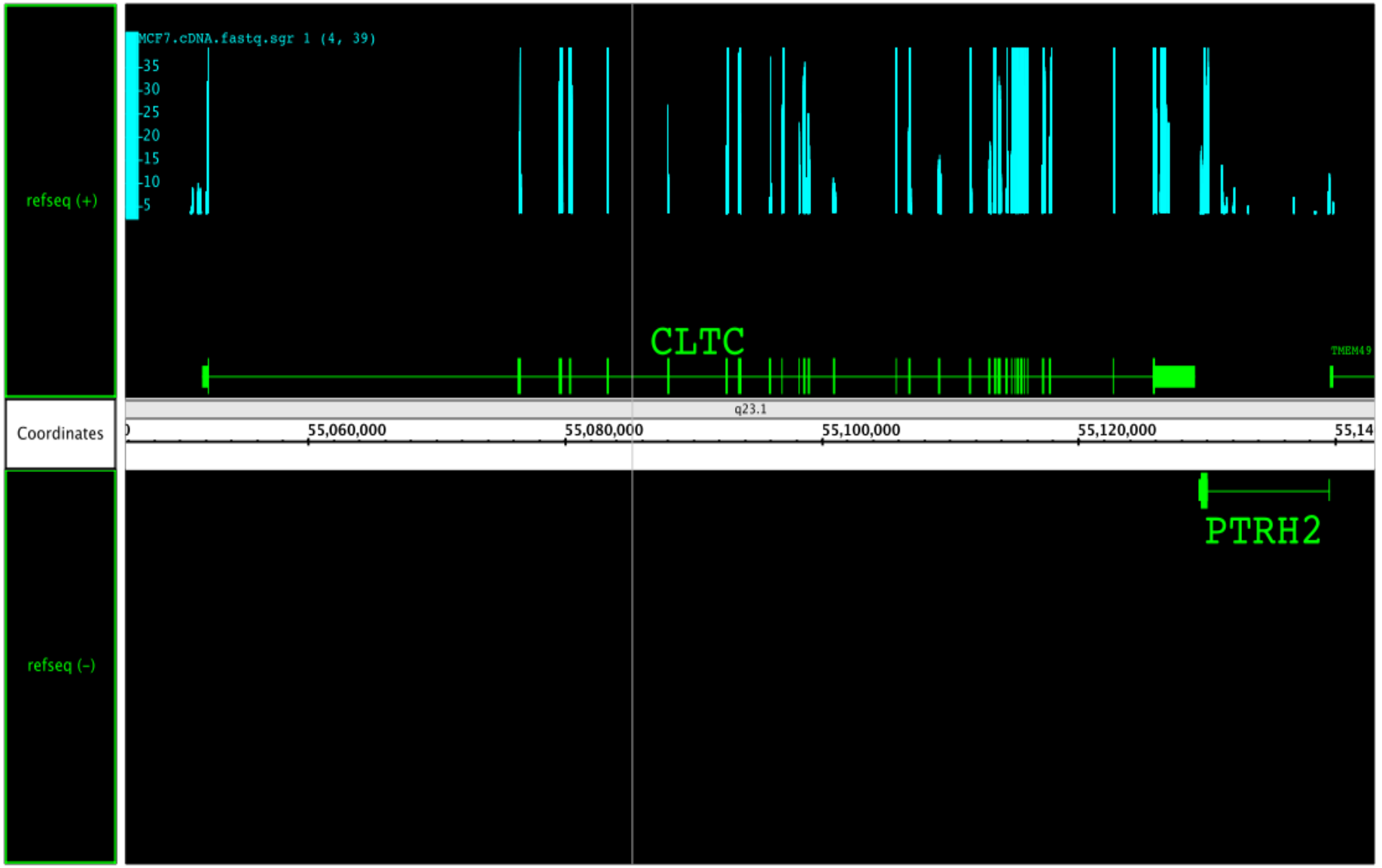
Repeats


chr20:
49584400
49584500
49584600
49584700
49584800
49584900
49585000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BAC End Pairs
User Supplied Track
User Track
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Repeats are not created equal

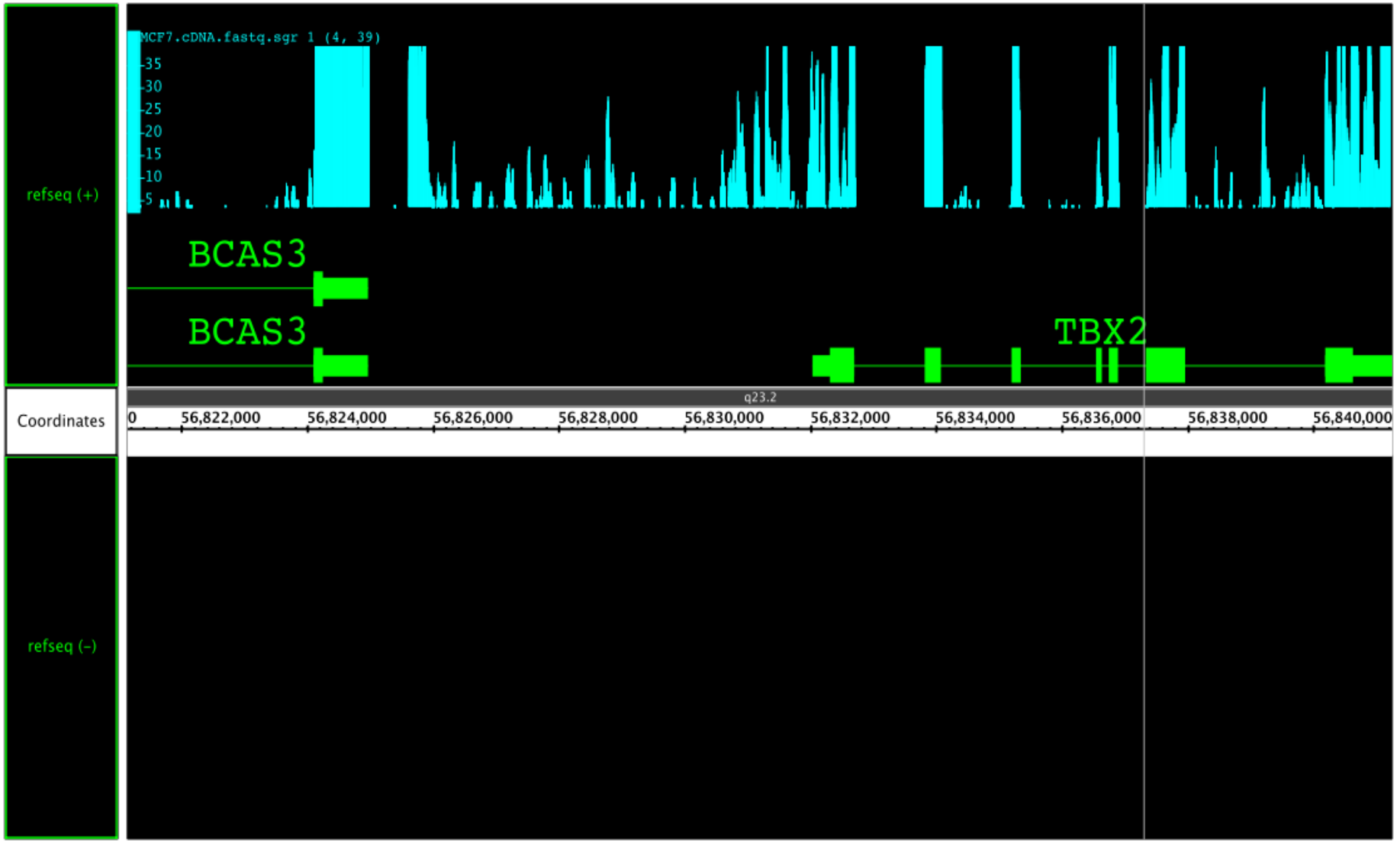


Transcription of Exons

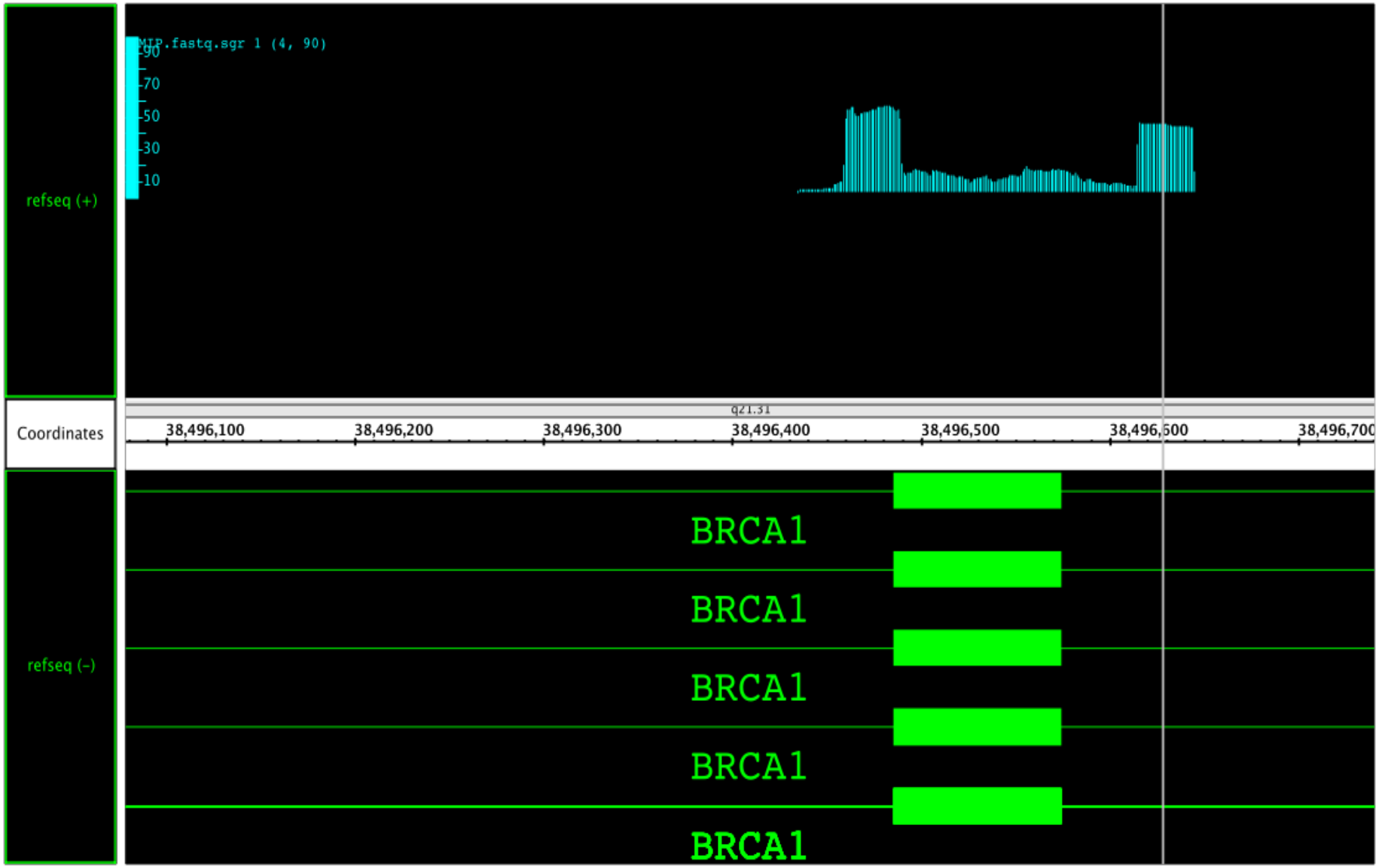


Lack of Transcription

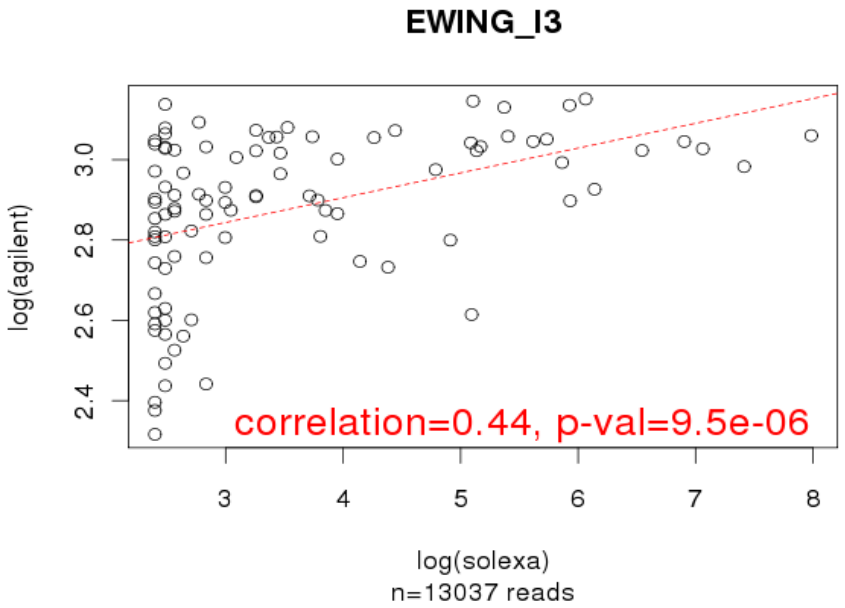
Transcription

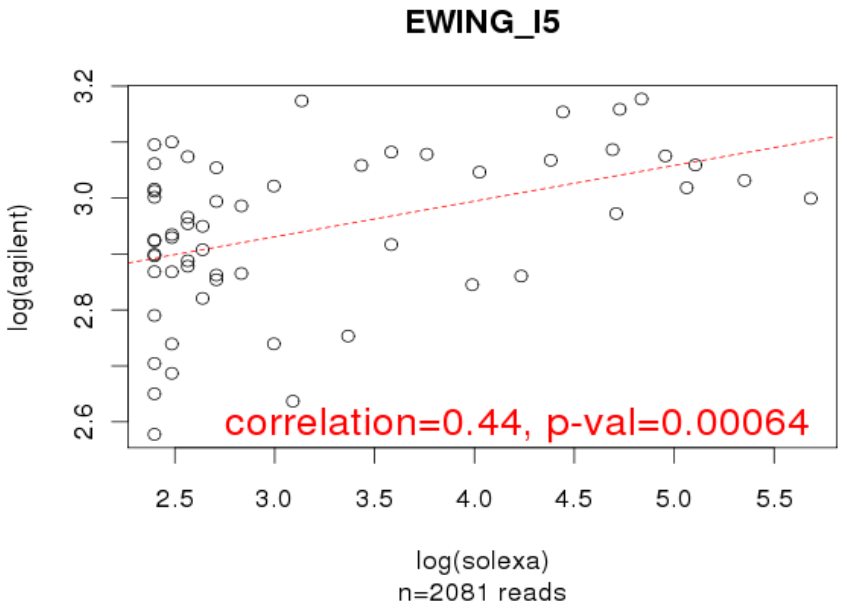


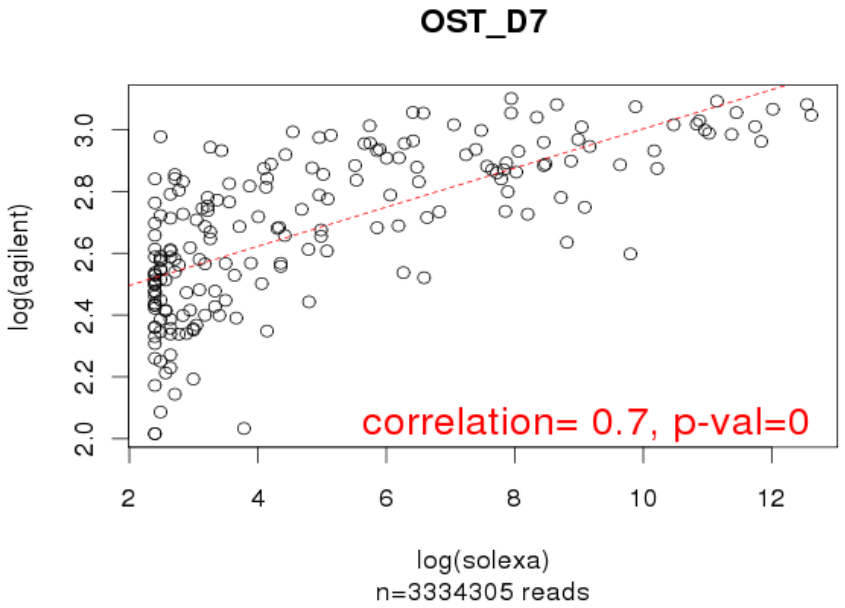
Targeted Sequencing

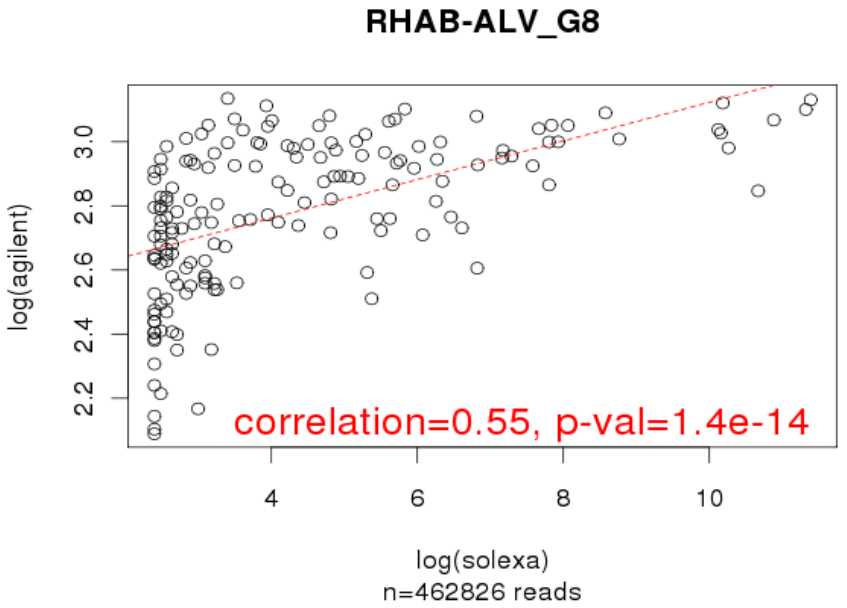


Sarcoma miRNA









The Data

- Raw data
  - What is it?
- The pipeline
  - Firecrest
  - Bustard
  - GERALD

Software

- Next-gen sequence analysis software
- Image Analysis/Base Calling
  - SwiftNG
  - Alta-cyclic
  - Rolexa
- Alignment
  - MAQ/bwa
  - Bowtie
  - Eland

Software

- Peakfinding
- SNP-calling
  - MAQ
  - SAM TOOLS

Software

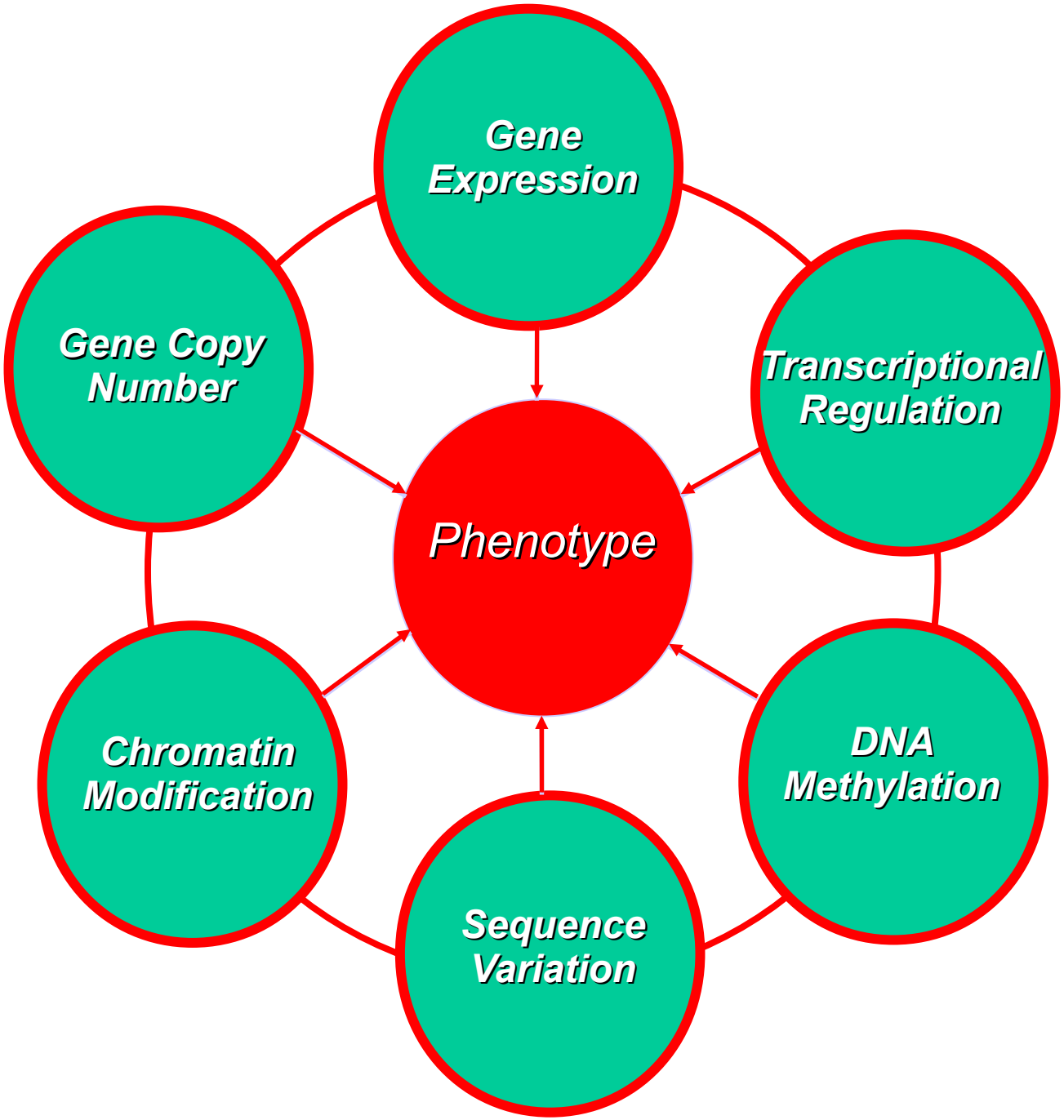
- Visualization/Data integration
  - Affymetrix Integrated Genome Browser
    - Extensions by Yevgeniy
  - Integrated Genome Viewer (Broad)
  - UCSC Genome Browser
  - Galaxy, Penn State University
  - R

Public Data

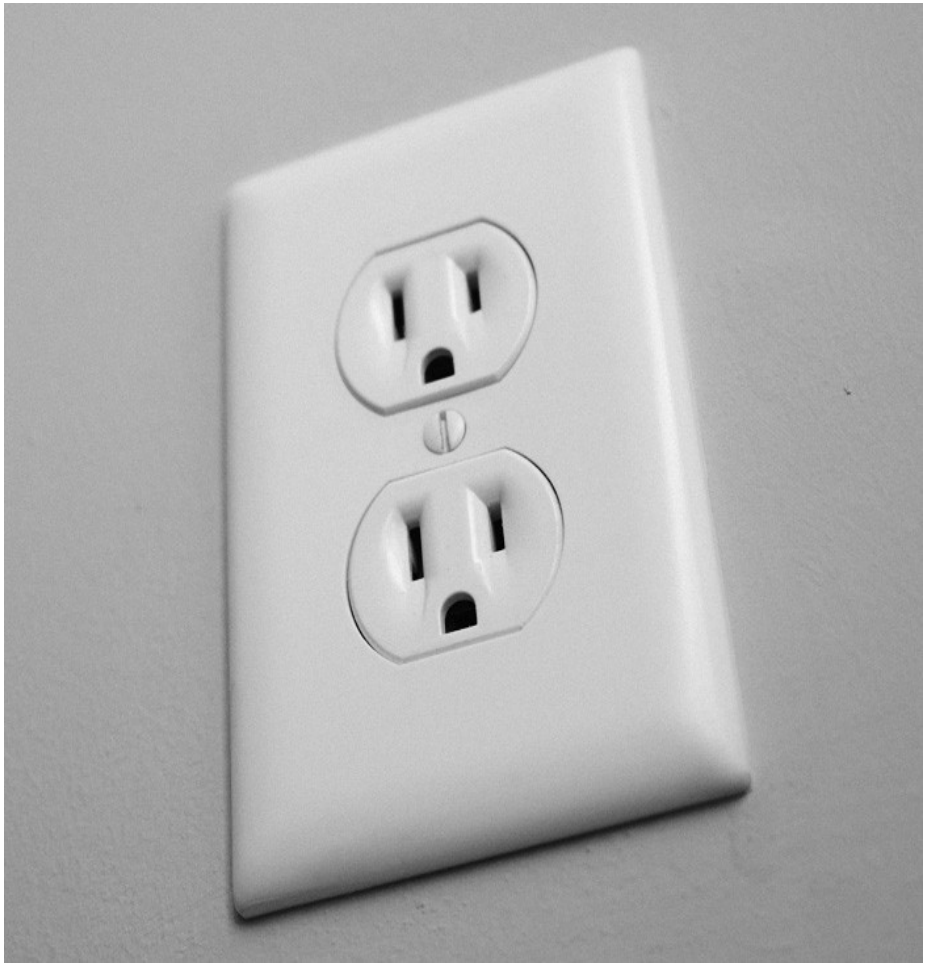
- SRA (NCBI)
- SRA (EBI)
- ....

Annotation and Data Management

- Plan for later submission to public repositories
- Keep public data available
- Keep track of generated data
- Reduce the number of copies of data passed around


Gene Expression
Gene Copy Number
Transcriptional Regulation
Phenotype
Chromatin Modification
DNA Methylation
Sequence Variation

The Last Mile



Thanks

- NCI
  - Paul Meltzer
  - Sven Bilke
  - Yuan Jiang
  - Bob Walker
  - Jack Zhu
  - Yevgeniy Gindin
  - Keith Killian
  - Princy Francis